网络行为数据的适用性评估问题初探*

■ 聂磊　王延飞
北京大学信息管理系　北京 100871

摘要：[目的/意义]探讨网络行为数据适用性的意涵、影响因素和评估方法，为相关研究提供参考，以促进此类数据的科学使用。[方法/过程]利用文献法梳理出网络行为数据适用性的核心影响因素，进而以此为基础，结合情报素材评估和社会调查数据评估领域的已有成果，探索如何对网络行为数据的适用性进行评估。[结果/结论]最终提出符合网络行为数据特征的适用性评估框架与方法，并结合案例初步验证所提方法的可用性。

关键词：网络行为数据　数据适用性　代表性　效度　情报感知
分类号：G250
DOI：10.13266/j.issn.0252-3116.2019.06.004

1 引言

获取足够的样本是社会科学量化研究的基础保障，但受制于现实条件，这在很多研究中难以实现[1]。大数据的兴起为这一问题的解决提供了新契机，研究人员能够以前所未有的水平观察社会现象，开展更多样、深入的研究[2]。作为大数据的典型代表，与搜索行为、社交媒体言论、电商平台消费等相关的网络行为数据在经济社会问题研究中正得到越来越多的使用，但是这些数据本身在特定的研究任务语境下是否适用或是否会对研究结果致偏等尚不明朗。

情报感知是情报工作的核心业务，要求情报专业人员清晰地认知、解读和表达情报用户需求、情报对象内容以及情报任务组织[3]。近年来，网络行为数据作为开源情报的重要来源之一，在新时代背景下的情报工作中正扮演着日益重要的角色[4]。而这些复杂、多样的网络行为数据的引入给情报感知中的数据感知提出了新的难题，例如这些数据是否可以用于满足情报用户的特定需求等。

本文将在总结已有研究的基础上探讨网络行为数据在涉及经济社会议题的研究与应用中的适用性问题，探析网络行为数据适用性的意涵、核心影响因素以及评估框架与方法。

2 网络行为数据适用性的意涵与研究现状

本文用“数据适用性”表征数据与研究问题的契合程度，即特定的数据在多大程度上适合解答特定的研究问题。对于这一概念所包含的具体内容，目前尚无统一界定，本文参考社会调查数据评估的已有研究，围绕以下两个方面讨论网络行为数据的适用性：①样本代表性，即数据中的样本能够被据以有效推估总体的程度；②测量效度，即基于特定数据进行的测量能在多大程度上反映研究概念的真实含义[5]。

严谨的研究离不开有效、高质量的测量，但这在网络行为数据及其所代表的大数据的应用中常难以实现[6]。多数常用大数据并不是为了科学研究而生[7]，其生产和采集过程中常缺少对数据有效性的提前设计[8]。一旦离开适用的情境，大数据将失去意义和价值[9]。J. Merino 等指出，在应用过程中，适用性（adequacy）是评估大数据质量的主要标准[10]。欧委会大数据质量任务团队指出大数据对目标总体的代表性是重要的质量问题[11]。J. Liu 等指出，大数据，尤其是互联网大数据，在为研究带来可能性的同时也带来了5个“大误差”，其中就包括数据源的非权威性和数据的代表性问题[12]。Z. Tufekci 着重讨论了社交媒体数据的

* 本文系国家社会科学基金重大项目“情报学学科建设与情报工作未来发展路经研究”（项目编号：17ZDA291）研究成果之一。
作者简介：聂磊（ORCID：0000-0003-1995-4114），博士研究生；王延飞（ORCID：0000-0002-2627-8153），教授，博士生导师，博士，通讯作者，E-mail：yfwang@pku.edu.cn。
收稿日期：2018-08-30　修回日期：2018-11-05　本文起止页码：29-34　本文责任编辑：易飞

适用性问题，指出社交媒体数据的代表性和效度将对研究的有效性带来巨大挑战[13]。

面对不完美的数据，研究者需要意识到问题的存在并用科学的方法降低其对研究的影响。对于如何解决网络行为数据或大数据的适用性问题，目前已有一些讨论。J. Merino 等提出了大数据适用性的测量框架和流程，但没有提出如何进行每一项测量[10]。D. Lazer 等建议研究者要了解影响大数据产生的算法，并将大数据和“小数据”结合使用，以减少研究中的偏误[7]。Z. Tufekci 指出可以通过定位非社会性因变量、重抽样、开展基线测量、与产业界合作等方法应对社交媒体数据的代表性和效度不足问题[13]。J. Liu 指出研究者需要更好地理解和评估大数据带来的“大误差”，与数据提供者合作以严谨的方法采集数据、用传统数据对大数据形成补充、扩展多源数据等方法有助于减轻这些误差带来的副作用[12]。黄恒君等提出通过单来源渠道权威性评价、数据生成机制分析、技术检查、替代型数据源印证、互补型数据源印证、信息可用性筛选等方法对网络商户数据进行评估[14]。

由此可见，网络行为数据及其所代表的大数据的适用性问题可能在研究和应用中导致严重的后果，这已经引起学界的高度关注。关于如何应对这一问题，现有讨论多是方向性的，操作层面的系统研究仍旧不足。就操作层面而言，情报学中的情报素材评估[15-16]和社会学中的调查数据评估[5,17]都能为网络行为数据的适用性评估提供参考依据，然而由于网络行为数据具有数量大、更新频率快等特点，现有评估方法并不直接适用。本文将着眼于网络行为数据的特点考虑数据适用性问题的核心影响因素，在现有评估方法的基础上，探索如何在操作层面评估网络行为数据的适用性，使网络行为数据能够更好地服务于社会科学研究与情报工作。

3 网络行为数据适用性的核心影响因素

网络行为数据的适用性问题源于其生产或产生机制。网络行为数据并非为统计而生，而是由有机系统根据业务需要记录下来的[18]，决定数据内容和形式的常常不再是研究人员，而是在多数情况下完全独立于研究人员的系统平台。不仅如此，“网络”一词天然地将行为发出者限定为网络用户。平台和用户是影响网络行为数据适用性的主要因素。

3.1 平台因素

目前最常被使用的网络行为数据，如社交媒体数据、搜索数据等，大多是由特定平台采集、存储并展示的，并且这些平台多具有商业属性，这就会从以下 3 个方面影响数据的适用性：

3.1.1 平台和用户的双向选择会影响数据代表性 网络行为数据并非全样本数据，以新浪微博为例，其 2017 年 12 月的月活跃用户数为 3.92 亿[19]，体量巨大，但也仅占中国总人口的 28.2%。数据量的增加并不能保证数据代表性[9]，尤其是来自商业平台的数据[12]。用户和平台之间的双向选择，如平台间差异化的营销策略、用户偏好等，会导致不同平台可能代表不同用户群体，因此一旦研究对象总体超出了平台范围，特定平台数据的适用性就成了问题。值得注意的是，这并不仅是抽样问题，更是机制(mechanisms)问题，在单一平台进行抽样并不能解决这一问题[13]。

3.1.2 平台会影响用户行为，进而影响数据测量效度 平台常常采用各种方法吸引和留存用户，如各类平台的精准营销、搜索引擎的关键词补全和推荐、社交媒体的热门事件推送等，这些都会对用户行为产生影响，导致行为数据背后的含义发生变化，进而导致看似有效、甚至曾经有效的测量变得无效，谷歌流感预测就是典型案例。D. Lazer 等指出工程师优化服务和用户接受服务的过程会产生算法动态性问题(algorithm dynamics)，即用户的行为随算法发生变化，而这正是导致谷歌流感预测无法持续成功的主要原因之一[7]。

3.1.3 平台对数据的管理行为会影响研究人员对数据的获取 网络行为数据作为大数据的典型代表，具有体量大、更新速度快等特征[20]，针对这些特征，平台在数据的记录、存储、检索等方面有相应的管理模式，包括元数据记录的规范性和完整性、数据入库周期、可供不同用户检索的数据范围、数据接口的调用范围等。这些模式由平台决定，通常不对外公开且随时可能变化，而其形成和变更常常是由商业目的决定的。这些模式必然影响研究者获取数据的数量、形式、甚至内容，进而影响数据的适用性。例如元数据中对数据属性的不规范记录可能导致研究者对数据产生错误理解进而提取出不适用于特定研究的数据。

3.2 用户因素

网络行为数据是由网络用户产生的，但社会科学研究的对象并不仅仅包括网络用户，因此用户因素对数据适用性的影响常常是不可避免的，其影响主要来源于以下 3 方面：

3.2.1 互联网用户的人口社会属性 即使不考虑平台因素，网络用户的总体特征也会影响数据的代表性，

即网络用户的人口社会属性可能与研究对象不一致。例如,截止到2017年12月,中国网民中城镇网民占比73%,农村网民占比为27%[21],而同期我国总人口中城镇人口占比58.52%,乡村人口占比为41.48%[22],两个分布差异较大,如果不加处理直接用网络行为数量对比的结果反映某一问题的城乡差异,就可能导致结果有偏差。

3.2.2 相同行为模式可能代表不同含义 一方面,不同、甚至相同类型的用户可能按照完全不同的逻辑实施同样的行为,各类行为数据混杂在一起可能使测量结果无效,进而导致其适用性大打折扣;另一方面,如果研究者获取数据的方式与用户行为模式不一致,也可能导致所得结果不适用,例如通过标签(hashtag)获取社交媒体数据就会使不喜欢加标签的用户被排除在测量之外[13]。

3.2.3 用户对"被测量"做出的反映 网络行为数据可被测量已不是秘密,如果用户不愿意被测量,就可能采取相应的策略使自己的行为"不可见"。例如G. Lotan的研究表明,部分推特用户在使用各种策略与推特做对抗,并在这一过程中很好地理解并利用了推特的倾向主题算法[23]。类似的做法在国内也很常见,例如通过图片、表情、暗语等表达观点,这都增加了测量的难度,使研究者难以获得适用于特定研究的数据,甚至可能在不自知的情况下获得不适用的数据。

4 网络行为数据适用性的评估框架与方法

通过以上分析不难发现,平台和用户对网络行为数据的适用性有着重要影响,因此可基于平台和用户进行评估。与此同时,还可以由后向前倒推,即对利用网络行为数据得到的测量结果进行评估,反过来判断数据的适用性。总体评估思路如图1所示:

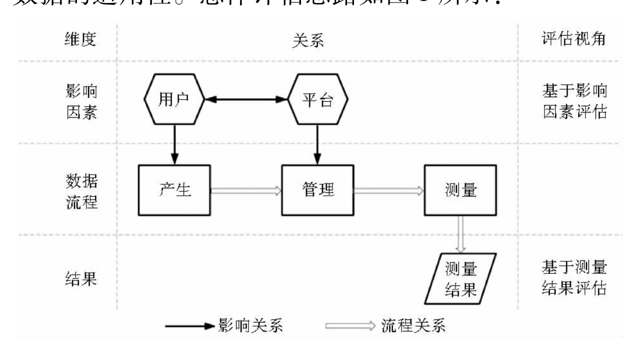


图1 网络行为数据适用性的评估思路

在这一思路下的评估框架如表1所示,其中基于影响因素评估对应的内容源于平台和用户对网络行为数据适用性的影响,基于测量结果评估对应的内容参考了社会调查数据的效度评估[17]。

表1 网络行为数据适用性评估框架

评估对象	评估视角	评估内容
网络行为数据适用性	基于影响因素评估	平台和用户特征
		平台对用户的影响
		平台和用户的行为模式
	基于测量结果评估	预测效度
		共变效度
		建构效度

4.1 基于影响因素评估

4.1.1 评估平台和用户特征 对于情报素材,评估其发生源的特征有助于判断其适用程度[15]。对于网络行为数据而言,其发生源由平台和用户共同组成,而通过上文讨论不难发现,互联网平台和用户都存在代表性问题,在利用网络行为数据研究总体时容易产生误差。但对研究者而言,重要的不是没有误差,而是能知道和控制误差的大小[17],尤其是当误差可能发生在研究的核心变量上时。对于数据代表性的评估,可参考如下评估方法:

(1)基于平台官方数据评估。大型平台常常会分析自己用户的特征并公开发布,其中包括相对精确的数据,如公司财务报告中的用户数量,也包括估算的数据,如用户画像。通过这些数据能对数据代表性形成方向性判断。

(2)基于已有调查数据或研究成果评估。一方面,可以利用以中国互联网络信息中心(CNNIC)系机构发布的数据评估数据代表性;另一方面,也可以通过已有实证研究获取判断依据,如M. Duggan和J. Brenner对推特的分析有助于研究者评估推特用户特征[24]。

以G. Doyle利用推特数据对英语方言演变进行的研究为例,作者首先基于已有研究指出推特数据偏向年轻群体,并且略微偏向城市,但语言学研究表明城市年轻人是语言变迁的主要驱动力,同时自媒体上语言的非正式性也符合语言变迁研究的要求,因此作者认为推特上的用户行为数据适用于这一研究[25]。

4.1.2 评估平台对用户的影响 关于平台对用户行为影响的研究目前相对较少,同时,这种影响可能是动态和不规则的,如平台会不定期推出新功能。因此,除了参考已有研究外,研究者更加需要通过实验评估具体研究案例中平台对用户的影响。

（1）单一平台的时序对比。其基本思想是，如测量对象具有一定的时序特征，则其时序数据的"异常"变化多是外力影响的结果。若平台行为已知，可评估其是否对表征用户行为的时序数据产生了影响；若平台行为未知，可通过时序数据"异常"识别进行辅助判断。例如可以通过差分值、时间序列分解后的随机因素值等指标在特定时间段内的方差判断其变异程度，或通过异常值识别算法发现数据异常。K. H. Bordersen 等的研究结果表明，通过时间序列模型能够验证某一条广告是否影响了谷歌用户的搜索和点击行为[26]，这一方法可推广至网络行为数据的适用性评估。

（2）多平台对比。这一方法的基本思想是，对于同一个问题的测量在多个平台得到的数据之间一致性越高，则单一平台对用户产生特定影响的可能性越低。需要注意的是，当不同平台间的测量结果一致性较高时，有可能是各平台对用户产生了相同的影响，因此所用平台越多、平台间差异越大，则对比结果越具参考价值。黄恒君等在研究利用网络数据构建单位名录库时，通过对比大众点评网和糯米网的商户信息验证了数据的真实性和全面性[14]，其本质是对比不同网站上的商家行为，因此其做法在评估平台对用户的影响时同样适用。

4.1.3 评估平台和用户的行为模式　研究者在评估平台对数据的管理机制和用户行为模式影响时面临的问题本质上是相同的，即需要建立平台和用户行为与研究之间的联系。对于这一问题，可参考如下评估方法：

（1）通过官方信息评估平台数据管理模式对适用性的影响。部分平台会在开发者平台、行业论坛等渠道公布自身技术信息，研究人员也可以向平台客服咨询相关技术信息。从笔者调研结果来看，虽然其中不乏一些有用信息，如数据接口的抽样比例、更新周期等，但通过这些渠道获取的信息常常不充分且不及时，或许这正是相关学者常建议研究人员跟数据提供者合作[12-13]的原因。

（2）基于网络行为模式的相关研究评估数据的适用性。网络行为模式是近年来一个热门研究主题，情报学、计算机科学、心理学等学科已在这一领域取得了大量研究成果，通过对已有研究的回顾将有助于研究者评估数据的适用性。例如，孙毅、吕本富等在利用搜索引擎数据研究消费者信心时，通过已有研究验证了网络搜索行为与消费者信心的关联，进而构建了基于搜索数据的消费者信心指数[27]。

（3）基于实验评估平台和用户的行为模式。如果缺乏已有研究和官方信息作为评估依据，研究者只能通过实验逆向研究用户行为模式和平台对数据的管理模式。虽然研究目的不同会导致实验方向的不同，但核心思想都是首先挖掘已经受到平台和用户行为影响的数据，发现其中的规律，进而逆向理解其产生过程。例如通过用户行为对用户进行聚类，并研究类别间行为模式的差异，进而探索不同行为模式所代表的含义，或者通过对数据的动态跟踪分析研究平台的数据发布规律。为说明这一思路，本文进行了一个简单的示例性实验。

笔者利用某社交媒体的站内搜索功能，以"研究"为关键词，区域限定为"北京"，采用相同检索条件，在不同时间点进行了多次检索，获得的数据条数如表 2 所示。检索的时间范围是 2018 年 5 月 16 日 8:00 到 11:59，从表中不难发现数据条数呈递减趋势，5 天时间内数据条数减少了 2.1%。虽然单次实验本身不能证明存在规律，但它提供了一种方向。假设这一结论经过大量实验验证，无论是由于部分用户倾向于删除行为数据，还是由于站内搜索引擎的限制，这都意味着研究者用这种方式获取的历史数据可能是不全面的，如果研究问题对这一点很敏感，尤其是研究发生在很久之前的事件时，这一数据的适用性就会大打折扣。

表 2　社交媒体平台站内搜索数据条数

检索时间	5.16 12:01	5.16 13:01	5.16 14:01	5.16 15:01	5.16 16:01
数据条数	1 125	1 125	1 124	1 124	1 123
检索时间	5.17 7:30	5.18 7:30	5.19 7:30	5.20 7:30	5.21 7:30
数据条数	1 116	1 111	1 107	1 102	1 101

4.2 基于测量结果评估

如果研究者无法对平台和用户进行评估，还可以通过评估测算结果倒推网络行为数据的适用性，这方面可借鉴社会调查数据的效度评估方法中预测效度、共变效度和建构效度评估。

4.2.1 评估预测效度　预测效度是"将已得到的测量结果与未来实际发生的情况进行比较，以检查两者的一致性。"[17] 当测量具有时序属性时，可采用预测效度进行评估。具体来看，有两种不同的方法：

（1）先测量，然后等待结果出现，最后进行评估。以统计指标的替代指标为例，多数统计指标的发布都会有所滞后，例如 6 月中旬发布对 5 月的测量结果，如果研究者在统计指标发布之前通过网络行为数据完成了测量，就可以在统计数据发布后对其进行验证，单次

测量可计算预测值与实际值的差值,多次测量可计算二者的均方误差等,进而通过这些指标判断数据的适用性。这是一种理想情况,但在实际使用中会受到一定限制。例如,若结果的发生时间是不确定的,则研究周期可能被无限拉长。因此在实际使用中有时会采用第二种方法。

(2)利用历史数据进行评估。其基本思想是,如果能用某一历史结果出现之前的数据精准地对其进行预测,则说明数据曾经是适用的,进而遵循时间序列外推的思路,认为数据现在仍有一定的适用性。如果历史数据序列足够长,则可使用均方误差、累计均方误差等指标进行评估。在实际使用中,研究者常用重大历史事件作为预测对象。这种方法易于实现,但属于事后解释,加上外推法本身的缺陷,其科学性容易受到质疑,因此常作为辅助评估方法。

基于网络行为数据的经济预测研究常使用预测效度评估数据的适用性。例如谷歌科学家 S. L. Scott 和经济学家 H. R. Varian 在利用谷歌趋势数据进行经济预测研究时,通过对比纯时间序列模型和加入谷歌趋势的时间序列模型,发现后者具有更小的预测误差,能够更好地预测 2008 年至 2009 年的经济危机,因此指出谷歌趋势数据在此类研究中具有重要价值[28]。

4.2.2 评估共变效度与建构效度 除了预测效果以外,还可以通过相关性评估其效度。其基本思想是如果测量的对象与已知对象在概念上相同或高度相关,那么当测量结果与表征已知对象的数据高度相关时,可以更有信心地认为测量是有效的,根据相关性的类型不同,评估依据可分为共变效度和建构效度。

(1)共变效度。共变效度用于判断新的测量能否取代现有测量[17],即用网络行为数据测量一个已知变量,如果新的测量结果与已知结果高度相关,如相关系数较大、回归系数显著等,则可以认为它是有效的。共变效度常用于现有测量认可度较高、但难度较大的情况,例如大型社会调查。如果研究人员要利用网络行为数据测量与现有指标相同或相似的概念,则可使用共变效度。

共变效度在实证研究中已得到使用。例如,G. Doyle 通过将 SeeTweet 测算结果与高质量但极其耗时的《北美英语地图集》和哈佛方言调查结果相比较,验证了 SeeTweet 在其方言研究中的适用性[25]。孙毅、吕本富等通过回归分析验证了基于搜索数据的网络通胀预期与消费者物价指数的相关性[29]。

(2)建构效度。建构效度评估是要了解"测量工具是否反映了概念与命题的内部结构"[17],其基本思想是当测量对象与另一对象在理论上高度相关时,如果在经验层能够证明二者相关,则更有信心认为测量是有效的。例如研究者利用网络行为数据测量了概念 A,其目的是为了研究概念 A 和 B 的关系,但不确定测量是否有效,此时如果已知概念 A 和概念 C 理论上高度相关,且概念 C 已经被量化,则可以通过 A 与 C 的回归系数的显著性检验对于 A 的测量是否有效,进而判断数据是否适用。

5 结语

网络行为数据在为社会科学研究和情报工作带来新机遇的同时,也带来了巨大的风险。如果研究者使用了不适用的网络行为数据而不自知,不仅可能导致结果有偏或无效,还可能产生负面的经济、社会影响。因此,网络行为数据的适用性问题研究意义重大。

本研究探讨了网络行为数据适用性问题的核心影响因素,并以此为基础,结合情报学和社会学已有的研究成果,提出一套科学、系统、可操作的网络行为数据适用性评估框架,填充了具体的评估方法,并结合案例初步验证了其可用性,以期为相关研究提供参考依据。

本研究仍处于初期探索阶段,后续仍需通过大量的文献研究和实验对现阶段的结论进行不断验证和扩充。这一过程中必然存在许多理论和方法论上的困难,例如利用网络行为数据验证其自身的适用性是否科学等,但随着这些问题的解决,网络行为数据及其代表的互联网大数据将会为社会科学的发展带来更多贡献。

参考文献:

[1] 金,基欧汉,维巴. 社会科学中的研究设计[M]. 陈硕,译. 上海:格致出版社, 2014.
[2] GOLDER S A,MACY M W. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures[J]. Science, 2011, 333(6051): 1878－1881.
[3] 王延飞,赵柯然,陈美华,等. 情报感知的研究解析[J]. 情报理论与实践, 2018, 41(8): 1－4.
[4] WILLIAMS H J,BLUM I. Defining second generation open source intelligence (OSINT) for the defense enterprise[EB/OL]. [2018－06－20]. https://www.rand.org/pubs/research_reports/RR1964.html.
[5] 巴比. 社会研究方法 [M]. 邱泽奇,译. 11 版. 北京:华夏出版社, 2009.
[6] POWER D J. Using 'big data' for analytics and decision support [J]. Journal of decision systems, 2014, 23(2): 222－228.
[7] LAZER D,KENNEDY R,KING G, et al. The parable of google flu: traps in big data analysis[J]. Science, 2014, 343(6176):

1203 – 1205.
[8] KAISLER S, ARMOUR F, ESPINOSA J A, et al. Big data: issues and challenges moving forward[C]//Proceedings of 2013 46th Hawaii international conference on system sciences. HI: IEEE, 2013: 995 – 1004.
[9] BOYD D, CRAWFORD K. Critical questions for big data[J]. Information communication & society, 2012, 15(5): 662 – 679.
[10] MERINO J, CABALLERO I, RIVAS B, et al. A data quality in use model for big data[J]. Future generation computer systems, 2016, 63(C): 123 – 130.
[11] UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE BIG DATA QUALITY TASK TEAM. A suggested framework for the quality of big data[EB/OL]. [2018 – 05 – 21]. http://www1.unece.org/stat/platform/download/attachments/108102944/Big% 20Data% 20Quality%20Framework%20-%20final-%20Jan08-2015.pdf? version=1&modificationDate=1420725063663&api=v2.
[12] LIU J, LI J, LI W, et al. Rethinking big data: a review on the data quality and usage issues[J]. Isprs journal of photogrammetry and remote sensing, 2016, 115(C): 134 – 142.
[13] TUFEKCI Z. Big questions for social media big data: representativeness, validity and other methodological pitfalls[C]// Proceedings of the eighth international AAAI conference on weblogs and social media. MI: AAAI, 2014.
[14] 黄恒君,陶然,傅德印. 单位名录库更新:互联网大数据源及其数据质量评估[J]. 统计研究, 2017, 34(1): 12 – 22.
[15] 王延飞,秦铁辉. 信息分析与决策 [M]. 2 版. 北京: 北京大学出版社, 2010.
[16] 周军. 论社会科学情报素材鉴别[J]. 情报资料工作, 2005(3): 46 – 48.
[17] 袁方. 社会研究方法教程[M]. 北京: 北京大学出版社, 2004.
[18] GROVES R M. Three eras of survey research[J]. Public opinion quarterly, 2011, 75(5): 861 – 871.
[19] 新浪科技. 新浪发布 2017 年第四季度及全年未经审计财报[EB/OL]. [2018 – 05 – 21]. http://tech.sina.com.cn/i/2018-02-13/doc-ifyrpeie3108640.shtml.
[20] GARTNER INC. What is big data? [EB/OL]. [2018 – 05 – 21]. https://www.gartner.com/it-glossary/big-data.
[21] 中国互联网络信息中心. 第 41 次中国互联网络发展状况统计报告[EB/OL]. [2018 – 05 – 21]. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/P020180305409870339136.pdf.
[22] 中华人民共和国国家统计局. 中华人民共和国 2017 年国民经济和社会发展统计公报[EB/OL]. [2018 – 05 – 28]. http://www.stats.gov.cn/tjsj/zxfb/201802/t20180228_1585631.html.
[23] LOTAN G. Data reveals that "occupying" twitter trending topics is harder than it looks[EB/OL]. [2018 – 05 – 21]. http://blog.socialflow.com/post/7120244374/data-reveals-that-occupyingtwitter-trending-topics-is-harder-than-it-looks.
[24] DUGGAN M, BRENNER J. The demographics of social media users – 2012[EB/OL]. [2018 – 05 – 21]. http://www.pewinternet.org/2013/02/14/the-demographics-of-social-media-users-2012.
[25] DOYLE G. Mapping dialectal variation by querying social media [C]//Proceedings of the 14th conference of the European chapter of the Association for Computational Linguistics. Gothenburg: Association for Computational Linguistics, 2014: 98 – 106.
[26] BRODERSEN K H, GALLUSSER F, KOEHLER J, et al. Inferring causal impact using bayesian structural time-series models[J]. Annals of applied statistics, 2015, 9(1): 247 – 274.
[27] 孙毅,吕本富,陈航,等. 基于网络搜索行为的消费者信心指数构建及应用研究[J]. 管理评论, 2014, 26(10): 117 – 125.
[28] SCOTT S L, VARIAN H R. Predicting the present with bayesian structural time series[J]. Social science electronic publishing, 2012, 5(1): 4 – 23.
[29] 孙毅,吕本富,陈航,等. 大数据视角的通胀预期测度与应用研究[J]. 管理世界, 2014(4): 171 – 172.

作者贡献说明:
聂磊:论文撰写;
王延飞:论文修改。

A Preliminary Study About the Adequacy Evaluation of Internet Behavior Data

Nie Lei Wang Yanfei
Department of Information Management, Peking University, Beijing 100871

Abstract: [**Purpose/significance**] This paper explores the meaning, influencing factors and assessment methods of the adequacy of Internet behavior data, in order to provide reference for related research and promote the scientific use of such data. [**Method/process**] This paper firstly uses the literature method to sort out the core influencing factors of the adequacy of Internet behavior data, and then based on this, uses the existing results in the field of information material evaluation and social survey data evaluation to explore how to evaluate the adequacy of Internet behavior data. [**Result/conclusion**] Finally, the evaluation framework and method of Internet behavior data are proposed, and the usability of the proposed method has been verified by research cases.

Keywords: Internet behavior data data adequacy representativeness validity information awareness